AD-A068 392 SYSTEMS CONTROL INC PALO ALTO CALIF F/G 12/2
NETWORKS OF QUEUES - APPROXIMATION RESULTS.(U)
MAY 77 A J LEMOINE N00014-76-C-0919
UNCLASSIFIED TR-19-2 NL

1 OF 1
AD A068392

END
DATE FILMED
6-79
DDC

SYSTEMS CONTROL, INC.
1801 PAGE MILL ROAD
PALO ALTO, CALIFORNIA 94304

TELEX: 348439

TELEPHONE (415) 494-1163

LEVEL II

SYSTEMS CONTROL, INC.
1801 PAGE MILL ROAD
PALO ALTO, CALIFORNIA 94304

TELEX: 348433

TELEPHONE (415) 494-1165

(1)

Technical Report No. 19-2

Final

May 16, 1977

1473

NETWORKS OF QUEUES - APPROXIMATION RESULTS

By

Austin J. Lemoine

Prepared Under Contract:

N00014-76-C-0919 (NR 042-365)

For the Office of Naval Research

Approved by:

E. Arlin Torbett

E. Arlin Torbett, Project Director
Manager Telecommunication Sciences Program
Systems Control, Inc.

CONTROL ANALYSIS CORPORATION
800 Welch Road
Palo Alto, California 94304

TABLE OF CONTENTS

ABSTRACT

This report is a sequel to our earlier review paper Lemoine (1977) on the equilibrium analysis of networks of queues. In this report we review some approximation results for networks of queues. Detailed discussion is limited to results which can be rigorously justified. In addition, we call attention to some important open problems. The bibliography includes references for results derived in a heuristic or informal fashion and includes background material on diffusion processes.

1.0 INTRODUCTION

In a previous paper Lemoine (1977) we provided an overview of available results on the equilibrium analysis of networks of queues and we called attention to some important open problems in the area. In this paper we review some approximation results for networks of queues. Detailed discussions are limited to results which can be rigorously justified. We do mention, however, the principal references for results derived in a heuristic or informal fashion. We also discuss some important open problems in the area of approximation results for networks of queues. Most of the results we describe involve Brownian motion and diffusion processes; a diffusion process is a strong Markov process with continuous sample paths. Consequently, the bibliography also includes a number of references on Brownian motion, diffusion processes and related topics.

Approximation results are available for both <u>open</u> networks and <u>closed</u> networks. In an open network customers originate from external sources and each customer eventually leaves the system. In a closed network a fixed and finite number of customers circulate through the network and no arrivals or departures are permitted. For open networks there are two main classes of results, both derived under the assumption that the "traffic intensity" of the network is equal to or approaching the value one. The first such class of results provides approximations, via diffusion processes with boundaries, for the time-dependent behavior of the vector of queue-length processes at the various nodes for acyclic networks and for a generalized version of the network model introduced in Jackson (1957); in an acyclic network a customer can visit a node only once. The second class of results for open systems

provides approximations, again via diffusions with boundaries, for the equilibrium distribution of the total waiting time experienced by an arbitrary customer in a general tandem queueing system whose "traffic intensity" is approaching the value one from below. For closed networks of queues with exponentially distributed service times at each node, approximation results are available for the equilibrium distribution of the queue-lengths vector, when the number of customers in the system and the number of servers at each node are large; in this same setting, some approximations for time-dependent behavior, via diffusion processes without boundaries, have been derived. The results for open systems and for closed systems are developed under the assumption that basic system parameters such as arrival rates and service rates do not vary with time.

The approximation results to which we have just referred are recent, and some have yet to appear in print. Moreover, many of the results are quite complicated. Thus, in order to convey the nature of these results in a straightforward manner, detailed presentations will be confined to relatively simple models, with indications given for more general situations.

This paper is organized as follows. In Section 2 we present a heavy traffic result for the time-dependent behavior of the multi-server queue due to Iglehart and Whitt (1970a,b). This result is of interest for networks because it also holds under fairly weak conditions on the customer arrival process and it can be generalized to provide joint limits for several facilities in an acyclic network in heavy traffic. This is followed by a discussion of the work of Harrison (1973b), (1977) on approximations for the waiting time process for queues in tandem, and the forthcoming work of Reiman (1977) on heavy traffic

results for a generalization of the network model introduced in Jackson (1957). In Section 3 we present some results for closed Markovian networks and we use a generalization of the classical repairman problem to illustrate those results. The results for the closed system include approximations for the equilibrium distribution due to Iglehart and Lemoine (1973), (1974) and Lureau (1974), and a diffusion approximation for time-dependent behavior due to Iglehart and Lalchandani (1973). The diffusion results of Iglehart and Lalchandani (1973) were originally derived by heuristic arguments; however, using methods to be reported in the forthcoming monograph of Varadhan (1977), these results can now be rigorously justified. Finally, in Section 4 we discuss some important open problems in the area of approximations for networks of queues.

The literature on Brownian motion and diffusion processes, and the allied subject of stochastic integrals and stochastic differential equations, is quite varied in scope and in level of difficulty. Some references incude the following, which are listed in approximate order of increasing level: Cox and Miller (1965), Karlin and Taylor (1975), Arnold (1974), Breiman (1968), Varadhan (1968), Freedman (1971), Friedman (1975), (1976), Gikhman and Skorokhod (1969), Itô (1961), Dynkin (1965), Gikhman and Skorokhod (1972), Varadhan (1977), McKean (1969), and Itô and McKean (1965).

Finally, we close this section with comments about approximation results derived in a formal setting and results derived by informal approaches. The approximation results discussed in Section 2 and Section 3 are obtained as a by-product of rigorously proven theorems. However, while the rigorous justification of an approximation scheme is always desirable, actually proving that a certain diffusion "approximates" a given process of interest can be a very

difficult affair. These difficulties not withstanding, the significant advantage of tractability makes it very tempting to develop diffusion models as approximations for processes of interest, even without providing rigorous proofs. Programs of this sort have been carried out by several authors. The papers of Gaver (1968) and Newell (1968) and the monographs Newell (1971), (1973), (1975), (1977) are outstanding examples of this approach. Other interesting work in this regard includes McNeil and Schach (1973), McNeil (1973), Kobayashi (1974a,b), Gaver, Lehoczky and Perlas (1975), and Gaver and Lechoczky (1976a,b). These programs have been achieved by combining probabilistic intuition with a careful examination of the process one wants to approximate. In many instances the results obtained appear to be quite good and would certainly seem to justify this informal approach for analyzing complex systems under non-stationary conditions.

2.0 OPEN SYSTEMS

In this section we present some heavy traffic results for open systems. The main objectives of heavy traffic research are to describe unstable queueing systems and to approximate stable queueing systems. The term heavy traffic was introduced by Kingman (1961), (1962), (1965) and referred to queueing systems with a traffic intensity ρ less than but close to one; in this context, see also Köllerström (1974). In subsequent work the notion was expanded to include both single queueing systems with a traffic intensity $\rho \geq 1$ or sequences of queueing systems with traffic intensities $\{\rho_n\}$ such that $\rho_n \to 1$ as $n \to \infty$. A comprehensive account of heavy traffic research in queueing theory through 1974 can be obtained from Whitt (1968), Iglehart (1973) and Whitt (1974). The most important tool in heavy traffic research is the theory of weak convergence of probability measures on complete separable metric spaces; the main references on this subject are Billingsley (1968) and Parthasarathy (1967).

Our discussion of heavy traffic results in this section draws upon the work of Iglehart and Whitt (1970a,b), Harrison (1973b), (1977) and Reiman (1977). In Section 2.1 we discuss some work of Iglehart and Whitt (1970a,b) on multi-server systems. In Section 2.2 we discuss some work on single server queues in tandem due to Harrison (1973b), (1977), and some work of Reiman on a generalization of the classical open network model of Jackson (1957).

Other approximation results for open systems are developed in Harrison (1973a) and Crane (1971), (1973), (1974a,b). The paper of Harrison (1973a) considers a system in which a single assembler is making a piece of equipment consisting of K sub-assemblies, and each sub-assembly arrives at the assembler's

station according to a renewal process. This model was generalized in Crane (1971), (1974b) to allow more than one-assembler. In Crane (1973), (1974a) heavy traffic results are developed for some interesting network models which arise in various mass transit systems.

2.1 Some Heavy Traffic Results for a Multi-Server System

The multi-server queueing model considered here has the following structure. The customer arrival process is the superposition of r independent renewal streams and there are s independent servers, or service channels, with the arrival process and the service channels independent. Arriving customers join a single queue and are served in order of their arrival without defections. Interarrival times in the ith arrival channel (or renewal process) have mean $1/\lambda_i$ and finite variance α_i^2. Service times in the jth service channel are independent and have a common distribution with mean $1/\mu_j$ and finite variance β_j^2. Let $\lambda = \lambda_1 + \ldots + \lambda_r$ (total arrival rate) and $\mu = \mu_1 + \ldots + \mu_s$ (maximum service rate). The natural measure of congestion for this system is $\rho = \lambda/\mu$. We first discuss the case where $\rho = 1$ and

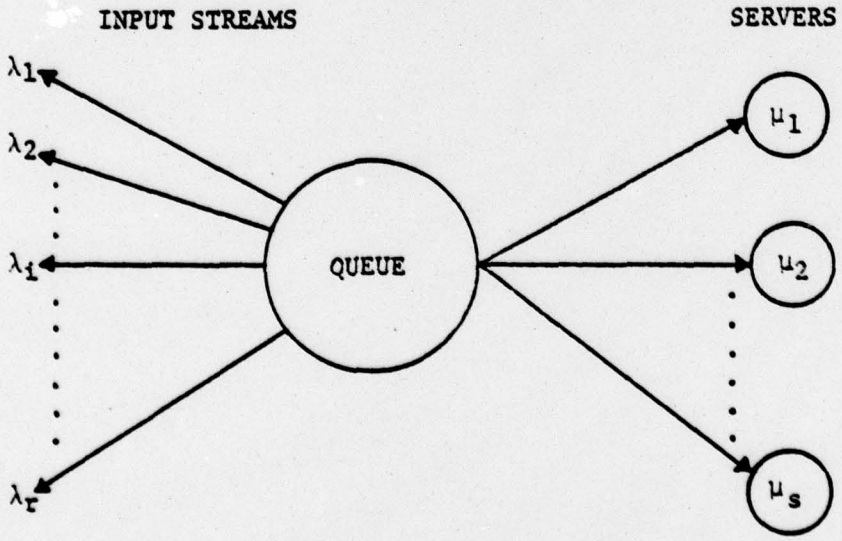

FIGURE 1 - MULTI-SERVER QUEUEING SYSTEM

then consider a sequence of such systems with traffic intensities $\{\rho_n\}$ such that $\rho_n \to 1$ as $n \to \infty$. In all cases we will be concerned with the time-dependent behavior of the stochastic process $\{Q(t),\ t \geq 0\}$ where $Q(t)$ is the total number of customers waiting for and receiving service at time t .

In the case where $\rho = 1$ the queueing system is unstable, with the variable $Q(t)$ becoming very large as time passes in the sense that $P\{Q(t) > k\} \to 1$ as $t \to \infty$ for any positive integer k . Just how fast this growth can occur is an important question. Unfortunately, the process $\{Q(t),\ t \geq 0\}$ is virtually impossible to work with in a direct manner. However, when $\rho = 1$ it turns out that the behavior of this process can be approximated by that of an appropriate diffusion process, through which the rate of growth can indeed be estimated.

The result we want to give is developed as follows. Let

$$\sigma^2 = \sum_{i=1}^{r} \lambda_i^3 \alpha_i^2 + \sum_{j=1}^{s} \mu_j^3 \beta_j^2 \ ,$$

and for $n = 1, 2, \ldots$ and τ in $[0, 1]$ let

$$Q_n(\tau) = \frac{Q(n\tau)}{\sigma n^{1/2}} \ .$$

For each n we are thus defining a random function on $[0,1]$ based on the sample paths of the process $Q(\cdot)$ over $[0,n]$. Let $\xi(\cdot)$ denote a standard Brownian motion process on $[0,1]$. Also, let $\Longrightarrow$ denote weak convergence

(or convergence in distribution); cf. Billingsley (1968). With this setup the following result is obtained in Iglehart and Whitt (1970a).

For all initial vlaues of $Q(0)$

$$Q_n(\cdot) \Longrightarrow |\xi(\cdot)| \tag{1}$$

as $n \to \infty$.

The limiting diffusion process in (1) is reflecting Brownian motion. From [26] there are the following corollaries to the above result, which provide precise estimates on the rate of growth of the process $Q(\cdot)$ as time passes. First, for $x \geq 0$

$$\lim_{t \to \infty} P\left\{\frac{Q(t)}{\sigma t^{1/2}} \leq x\right\} = 2\Phi(x) - 1 \tag{2}$$

where Φ is the distribution function of $N(0,1)$ random variable. Second, if

$$\hat{Q}_n = (1/\sigma n^{1/2}) \sup_{0 \leq t \leq n} Q(t)$$

then

$$\lim_{n \to \infty} P\left\{\hat{Q}_n \leq x\right\} = \Psi(x)$$

where

$$F(x) = 1 - (4/\pi)\sum_{k=1}^{\infty}\left[(-1)^k/(2k+1)\right]\exp\left\{-\left[\pi^2(2k+1)^2/8x^2\right]\right\} \quad .$$

Moreover, rates of convergence for (1) and other results from [26] and [27] are developed in Kennedy (1972a,b).

The result (1) can also be interpreted as follows. For n large the behavior of $Q(\cdot)$ over $[0, n]$ is, in a distribution sense, approximately that of $\{\sigma|n^{1/2}\xi(\tau)| \ , \ 0 \le \tau \le 1\}$; and, as $n \to \infty$ the process $n^{1/2}\xi(\cdot)$ goes to standard Brownian motion $\{B(t), 0 \le t < \infty\}$. Thus, the behavior of the process $\{Q(t), t \ge 0\}$ can be approximated by that of $\{\sigma|B(t)|, t \ge 0\}$. Hence, we would expect the process $Q(\cdot)$ to move in approximately the same direction and with approximately the same "speed" as the reflecting Brownian process $|B(\cdot)|$. Then, another device for estimating the rate of growth in the process $Q(\cdot)$ is to use the first-passage distributions of the process $|B(\cdot)|$.

So, for $z > 0$ let

$$T_z = \inf\{t : B(t) \ge z \text{ or } B(t) \le -z\} \quad ,$$

so that T_z is the epoch of first passage to the level z by the process $|B(\cdot)|$. The Laplace transform of the distribution of T_z is given by

$$E\left\{\exp[-\theta T_z] \,|B(0) = 0\right\} = 1/\cosh\left[z(2\theta)^{1/2}\right]$$

for $\theta \geq 0$; cf. Karlin and Taylor (1975). We then have

$$E\{T_z | B(t) = x\} = (z - x)^2$$

and

$$\mathrm{Var}\{T_z | B(t) = x\} = (z - x)^4/2$$

for all $0 < x < z$ and $t \geq 0$. For the first passage probabilities we have the following. For $0 < a < b$ and $t > 0$ let

$$L_b(t,a) = P\{T > t | B(0) = a\}$$

where

$$T = \inf\{s > 0 : B(s) = 0 \text{ or } B(s) = b\} .$$

Observe that

$$P\{T_z > t | B(0) = 0\} = L_{2z}(t,z)$$

for $z > 0$ and $t > 0$. It is known, cf. Feller (1971) or Freedman (1971), that

$$L_b(t,a) = \sum_{k=-\infty}^{+\infty} \left\{ \Phi\left([2kb + b - a]/t^{1/2} \right) - \Phi\left([2kb - a]/t^{1/2} \right) \right.$$

$$\left. + \Phi\left([2kb + b + a]/t^{1/2} \right) - \Phi\left([2kb + a]/t^{1/2} \right) \right\}$$

An alternate expression for $L_b(t,a)$ is given by

$$L_b(t,a) = (4/\pi) \sum_{k=0}^{\infty} [1/(2k+1)] \exp\left\{ - (2k+1)^2 \pi^2 t^2 / 2b^2 \right.$$

$$\left. \times \sin[(2k+1)\pi a/b] \right\} .$$

The first series converges rapidly for small t while the second converges rapidly for large t .

Returning now to the multi-server system, we see that if $Q(t) = m$ then the mean and the variance of the elapsed time until $m^* > m$ customers are in the system might be approximated by

$$(m^* - m)^2/\sigma^2 \quad \text{and} \quad (m^* - m)^4/2\sigma^4 \quad . \tag{3}$$

And, the probability of the customer population staying below the level m^* throughout the time interval $[t, t + t^*]$ might be approximated by

$$L_{\frac{2(m^* - m)}{\sigma}} \left(t^*, \frac{m^* - m}{\sigma} \right) .$$

A noteworthy feature of the estimates (2) and (3) is their extreme simplicity; only the basic system parameters λ_i , α_i^2 , $i = 1 \ldots, r$, and μ_j , β_j^2 , $j = 1, \ldots, s$, are needed for implementation.

An important direction in which the result (1) also holds is in specifying the limiting behavior of a sequence of multi-server queueing systems when the corresponding sequence of traffic intensities $\{\rho_n\}$, where $\rho_n = \lambda_n/\mu_n$, approaches the limit $\rho = 1$ from below; that is, we are interested in approximating the time-dependent behavior of systems which are stable but only just so. There will be different limits depending on the rate at which $\rho_n \uparrow 1$ as $n \to \infty$; the limiting behavior actually depends on $\lambda_n - \mu_n$ rather than λ_n/μ_n , but these are equivalent as long as $\lambda_n \to \lambda$ and $\mu_n \to \mu$ where $0 < \lambda$, $\mu < \infty$. To illustrate, let $\tilde{Q}_n(\cdot)$ denote the queue-length process for the nth system, and let σ_n^2 be defined for the nth system in the same way as σ^2 for the single system above. Let $Q_n^*(\tau) = \tilde{Q}_n(n\tau)/\sigma_n n^{1/2}$ for τ in $[0,1]$. For c in $(-\infty, 0]$ let $\xi^*(\tau) = \xi(\tau) + c\tau$, i.e., $\xi^*(\cdot)$ is Brownian motion on $[0, 1]$ with drift c . Now let

$$X^*(\tau) = \xi^*(\tau) - \inf_{0 \le \tau^* \le \tau} \xi^*(\tau^*)$$

Thus, the process $X^*(\cdot)$ is the process $\xi^*(\cdot)$ together with an impenetrable barrier at the origin. Then, the appropriate generalization of (1) is the following.

If $\sigma_n^2 \to \sigma^2$ **as** $n \to \infty$, **where** $0 < \sigma^2 < \infty$, **and**

$$n^{1/2}(\lambda_n - \mu_n) \to c \tag{4}$$

then as $n \to \infty$

$$Q_n^*(\cdot) \Longrightarrow X^*(\cdot) \quad . \tag{5}$$

If $c = 0$ then $X^*(\cdot)$ has the same distribution as $|\xi(\cdot)|$, cf. Itô and McKean (1965), and we obtain (1) as a special instance of (5). The more interesting case is where $c < 0$; in this case the result (5) says that the behavior of $\tilde{Q}_n(\cdot)$ can be approximated by that of Brownian motion with negative drift c/σ together with an impenetrable barrier at the origin. For more on this case see [27] and the references therein.

At first glance it might appear that the results (1) and (5) are not relevant for describing the behavior at a node (or service facility) in an open network of queues since the total customer input to a node seldom consists of the superposition of a finite number of independent renewal streams. However, the following weaker condition on the customer arrival process, say $\{A(t), t \geq 0\}$, is sufficient to insure that both (1) and (5) hold. For $n = 1, 2, \ldots$ and τ in $[0,1]$ let $A_n(\tau) = [A(n\tau) - n\tau\lambda^*]/\sigma^* n^{1/2}$ where λ^* and σ^* are specified positive constants. Then if

$$A_n(\cdot) \Longrightarrow \xi(\cdot) \tag{6}$$

as $n \to \infty$, both of the results (1) and (5) hold (with appropriate constants λ and σ). The "central tendency" condition (6) would be satisfied, for example, in the case where the process $A(\cdot)$ is the pooled output of a finite number of independent multi-server facilities in parallel, each of which is

is stable, or in the case where $A(\cdot)$ is the output of a finite number of multi-server facilities in tandem, each of which is also stable.

Significantly, the heavy traffic results (1) and (5) have been generalized in [27] to provide joint limits for several facilities in an acyclic network, such as the system depicted in Figure 2. A sequence of such networks is considered in [27] with the traffic intensity at each node in the nth system approaching the critical value 1 as $n \to \infty$. Multi-dimensional limits are obtained for the normalized vector of queue-length processes at the various nodes. However, the multi-dimensional processes which arise as

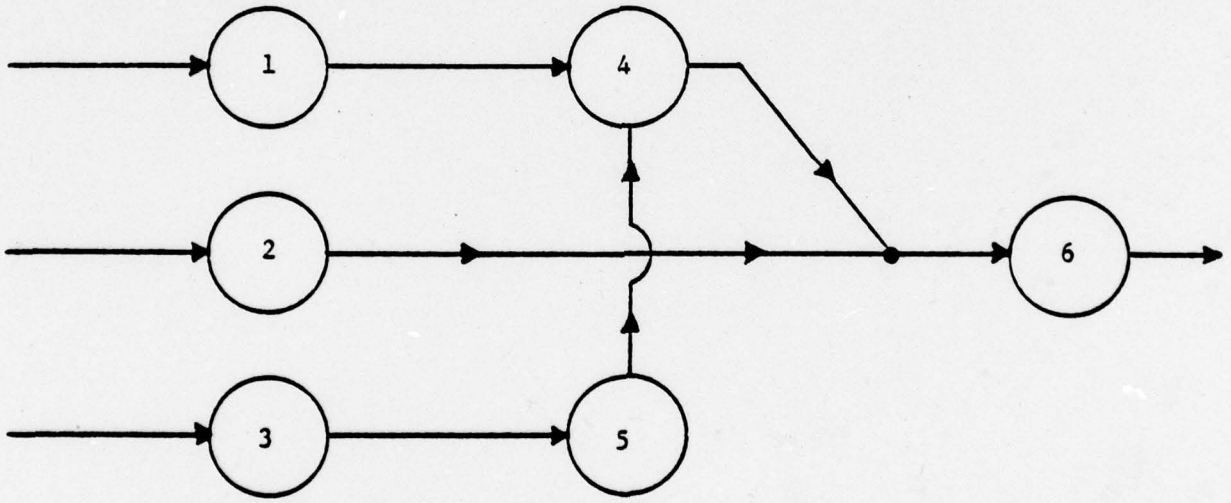

FIGURE 2 - ACYCLIC QUEUEING NETWORK

limits for these networks are very complicated functionals of multi-dimensional Brownian motion which are very difficult to evaluate in detail. As a consequence, there are as yet no computationally tractable approximations which have

been extracted from these joint limits, such as those given above for the multi-server system.

2.2 Some Results for Tandem Queues and for General Open Networks

We now discuss some work of Harrison (1973b), (1977) on tandem queues and some work of Reiman (1977) on a generalization of the classical network model of Jackson (1957). We encourage the reader to consult also the very interesting reports of Newell (1975), (1977); see also Kobayashi (1974a,b).

The tandem system consists of $K > 1$ single server facilities (or stations) arranged in series and a single external input process; see Figure 3. Customers arrive at the first facility according to a renewal process and proceed through the series of K stations. Customers receive service at each station in order of arrival, no customers defect at any stage, and all stations have unlimited queue capacity.

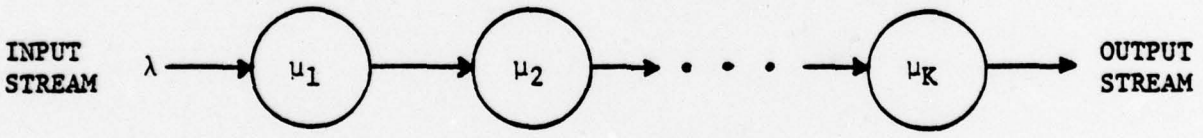

FIGURE 3 - SINGLE-SERVER QUEUES IN TANDEM

The interarrival times in the input process and the service times at each station are mutually independent sequences of independent and identically distributed random variables. The inter-arrival times are distributed as a variable U and the service times at station k $(1 \leq k \leq K)$ are distributed as a variable V_k. Let $E\{U\} = a$, $\mathrm{Var}\{U\} = \sigma_A^2$, $E\{V_k\} = b_k$ and $\mathrm{Var}\{V_k\} = \sigma_k^2$, $1 \leq k \leq K$, and assume that all of these quantities are finite. Let w_n^k denote the waiting time (exclusive of service time) at station k for the nth arriving customer, and let

$$w_n = \begin{pmatrix} w_n^1 \\ w_n^2 \\ \cdot \\ \cdot \\ \cdot \\ w_n^k \end{pmatrix} .$$

Now let

$$d = \min_{1 \leq k \leq K} (a - b_k) .$$

It is well known, and indeed quite clear, that w_n converges in distribution (as $n \to \infty$) to a proper limit w if and only if $\max_{1 \leq k \leq K} b_k/a < 1$, that is to say, if an only if $d > 0$. Thus, the parameter d can be regarded as the traffic intensity of the tandem system. In the case where $d > 0$ the system is said to be stable, but in what follows we say that the tandem system is in heavy traffic if d is positive but close to zero. The main result of

Harrison (1973b) provides a heavy traffic approximation for the equilibrium waiting time vector w, generalizing the result of Kingman (1961) for the queue GI/G/1. The main theorem of [23] shows that under heavy traffic conditions the vector dw is distributed approximately as a certain K-vector Z. The random vector Z is defined as a certain functional of $(K + 1)$ - dimensional Brownian motion, and its distribution depends upon the underlying interarrival and service time distributions only through their means and variances. The precise formulation of the main result in [23] is in terms of a limit theorem for a sequence of tandem systems with $d \downarrow 0$.

The methodology employed in Harrison (1973b) is similar to that of Iglehart and Whitt (1970a,b). There is, however, an important distinction between the results of [23] and the result obtained in [27]. As noted in Section 2.2, the paper [27] considers a sequence of acyclic networks with the traffic intensities at the various nodes approaching the critical value 1. The results of [27] are concerned with the time-dependent behavior of queue-length processes, showing weak convergence to certain diffusion processes related to Brownian motion. Such results, however, do not address the problem of heavy traffic behavior of equilibrium distributions, as do the results of [23].

We now present the main result of Harrison (1973b) as it related to the equilibrium waiting time vector w. The result is for a sequence of series queueing systems indexed by $m \geq 1$. Each system has K stations and satisfies all of the assumptions given above regarding the distributions of interarrival and service times. In what follows all of the notation introduced above will be maintained, except that a functional dependence on m will be added to indicate a random variable or a parameter associated with the mth system.

The result is built up as follows. It is first assumed that as $m \to \infty$

$$d(m) \downarrow 0 \quad ,$$

$$\sigma_A(m) \to \sigma_A \quad ,$$

$$\sigma_k(m) \to \sigma_k \quad , \quad 1 \leq k \leq K$$

and

$$[a(m) - b(m)]/d(m) \to c_k \quad , \quad 1 \leq k \leq K \quad ,$$

where each of σ_A , $\sigma_1, \ldots, \sigma_K$, $c_1, \ldots, c_K$ is non-negative and finite. (Note that necessarily $c_k \geq 1$ for $k = 1, \ldots, K$ as well). In addition, it is assumed that

$$E\left\{[U(m)]^{2+\epsilon}\right\} \quad \text{and} \quad E\left\{[V_k(m)]^{2+\epsilon}\right\}$$

are uniformly bounded in m and k for some $\epsilon > 0$. Now let $B^1(\cdot)$, $\ldots B^{K+1}(\cdot)$ be independent standard Brownian motion processes on $[0, \infty)$. For $1 \leq k \leq K$ and $t \geq 0$ let

$$C^k(t) = \sigma_k B^k(t) - \sigma_A B^{K+1}(t) - c_k t \quad .$$

Then, for $1 \leq k \leq K$ let

$$Y^k = \sup_{0 \leq z_k \leq \ldots \leq z_0 < \infty} \left\{ \sum_{j=1}^{k} \left[C^j(z_{j-1}) - C^j(z_j) \right] \right\}$$

and then put

$$Z = \begin{pmatrix} Y^1 \\ Y^2 - Y^1 \\ \cdot \\ \cdot \\ \cdot \\ Y^K - Y^{K-1} \end{pmatrix} . \qquad (7)$$

Observe that the random vector Z depends upon the underlying interarrival and service time distributions only through the parameters $\sigma_A, \sigma_1, \ldots, \sigma_K, c_1, \ldots, c_K$. Thus, we can write $Z = Z(\sigma_A, \sigma_1, \ldots, \sigma_K, c_1, \ldots, c_K)$. In the notation we have established, the random vector $w(m)$ is the equilibrium waiting time vector for the mth system. The result from Harrison (1973b) which we want to give can now be stated as follows.

<u>As $m \to \infty$ the random vector $w(m)$ converges in distribution to the random vector</u> $Z(\sigma_A, \sigma_1, \ldots, \sigma_K, c_1, \ldots, c_K)$.

The outstanding unsolved problem posed in [23] is the explicit determination of the distribution of Z for general values of the parameters σ_A,

$\sigma_1, \ldots, \sigma_K, c_1, \ldots, c_K$. It is demonstrated in [23], however, than in the special case where

$$0 < \sigma_A = \sigma_1 = \ldots = \sigma_K \equiv \sigma$$

the components of Z are independent, with the kth component $(1 \leq k \leq K)$ having an exponential distribution with parameter c_k/σ^2 . It is not likely, though, that Z will have such a simple distribution for arbitrary values of the constituent parameters. That this is indeed the situation is borne out by recent work on the tandem queue model of Figure 3 to be reported in Harrison (1977). We now describe some of this ongoing work of Harrison.

With w_n and d defined as above let

$$X(t) = d\, w_{[t/d^2]}$$

for $t \geq 0$. Now consider a sequence of tandem systems with $d \downarrow 0$ as above. What happens is that

$$X(\cdot) \Longrightarrow X(\cdot)$$

as $d \downarrow 0$, where $\{X(t), t \geq 0\}$ is a K-dimensional diffusion process defined (not surprisingly) as a complicated functional of multi-dimensional Brownian motion. The process $X(\cdot)$ is essentially the same as the process obtained by applying the results of Iglehart and Whitt (1970b) to the vector of queue-length processes at the various stations for the tandem system.

The first area of recent work on the tandem queue model to be reported in Harrison (1977) has been in the analytical characterization of the diffusion process $X(\cdot)$. Specifically, a good deal of effort has focused on the infinitesimal generator of the Markov process $X(\cdot)$, and this has a number of implications which we now discuss. (This work is not directly related to the main result of Harrison (1973b) since that paper deals exclusively with equilibrium distributions; but more will be said about that later on.) The state space of the process $X(\cdot)$ is, of course, the non-negative orthant in K dimensions; moreover, on the interior of the state space $X(\cdot)$ behaves as a K-dimensional Brownian motion with certain drift coefficients and a certain covariance matrix. The generator calculations which have been developed show that the behavior of $X(\cdot)$ at the boundary of the state space is instantaneous reflection at an angle which is constant along each boundary surface but different for each surface. For the case of two queues in series the angles of reflection are shown in Figure 4; the coordinates of the vector x

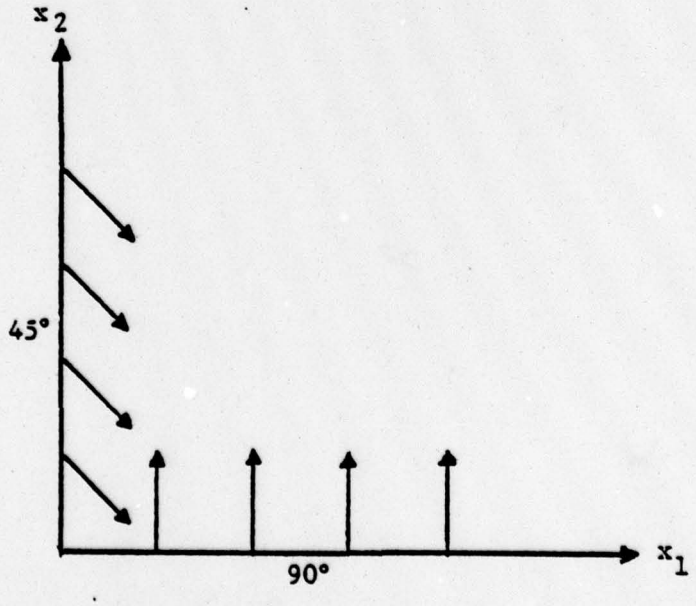

FIGURE 4 - ANGLES OF REFLECTION FOR DIFFUSION LIMIT OF WAITING TIME PROCESS FOR TWO SINGLE-SERVER QUEUES IN TANDEM

are denoted by x_1 and x_2, so that x_1 corresponds to the customer waiting time at the first station and x_2 to the waiting time at the second station. The generator calculations are also useful in the following context. For almost any distribution associated with the process $X(\cdot)$, such as occupation-time distributions, transient distributions, etc., the generator calculations suggest a corresponding partial differential equation; and if a solution to this partial differential equation can be found, or just shown to exist, then the generator results can be used to show that this solution is unique and coincides with the desired distribution.

Going back to the waiting time process in the tandem system, it has been shown that the diffusion $X(\cdot)$ has a stationary distribution $\hat{\pi}$, and that $\hat{\pi}$ is the distribution of the random vector Z given above as the weak limit of dw as $d \downarrow 0$. This means the following. Let $P_t(x,E) = P\{X(t) \in E | X(0) = x\}$ for Borel sets E in the non-negative orthant in K dimensions. Then

$$\hat{\pi}(E) = \int_E P_t(x,E)\hat{\pi}(dx) \quad ,$$

and for each K-dimensional vector y

$$P\{Z \leq y\} = \hat{\pi}(\{x : 0 \leq x \leq y\}) \quad .$$

More significantly, however, a partial differential equation (with boundary conditions) has been developed for the density of the stationary distribution $\hat{\pi}$. Specifically, it has been shown that if the K-dimensional probability

density function f satisfies this partial differential equation, then it is indeed the density function of the distribution $\hat{\pi}$, and hence the density for the correct heavy traffic approximation to the equilibrium waiting time distribution.

Thus, in order to completely resolve the outstanding problem posed in Harrison (1973b), it is necessary to produce a general solution for a certain partial differential equation and to prove that this solution is a probability density function. For the special case (given earlier) where $\sigma_A = \sigma_1 = \ldots = \sigma_K$, the corresponding f indeed solves the requisite partial differential equation. Another special case for which a solution has been found is the following. For $K = 2$ suppose that $\sigma_A = \sigma_2^2 \equiv \sigma^2 > 0$, $\sigma_1^2 = 0$, and $c_1 = c_2 \equiv c$. This corresponds to deterministic service times at station 1, the service time variance at station 2 equal to the interarrival time variance, and equal service rates at the two stations. For this case Z has density

$$f(x_1,x_2) = (x_1^2 + x_2^2)^{-1/4} \exp\left\{-c\left[x_1 + (x_1^2 + x_2^2)\right]\right\} \cos\left[\tan^{-1}(x_2/x_1)/2\right]$$

for $x_1 > 0$ and $x_2 > 0$. Observe that in this particular case the components of the random vector Z are not independent. As mentioned earlier, all of this work will be reported in Harrison (1977).

Finally we want to call attention to the forthcoming work of Reiman (1977) on heavy traffic approximations for open networks of queues. The model considered in Reiman (1977) is a generalization of the classical open Markovian network model of Jackson (1957). Specifically, there are N nodes with node i having a single server, a first-come-first served queue discipline, and a

waiting room of unlimited capacity. The external input stream to node i is a renewal process, the inter-arrival times in this process having mean λ_i^{-1} and finite variance. These external input streams at the various nodes are assumed to be independent. The service times at node i are independent and have a common distribution with mean μ_i^{-1} and finite variance. The service times at node i are also independent of the customer arrivals at node i. A customer leaving node i is immediately and independently routed to node j with probability p_{ij}; and the customer departs the system from node i with probability $q_i = 1 - \sum_{j=1}^{N} p_{ij}$. The matrix $\hat{P} = [p_{ij}]$ is called the switching matrix. Observe that this system is quite general, encompassing the tandem system, acyclic networks of $GI/G/1$ queues, and networks of $GI/G/1$ queues with feedback. The state of the network at time $t \geq 0$ is the N-vector $H(t)$ where the ith component of $H(t)$ is the total number of customers at node i at time t.

The notion of traffic intensity for this system is developed as follows. Let $\hat{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ and let $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2 \ldots, \hat{\alpha}_N)$ be the row vector solution of the equation

$$\hat{\alpha} = \hat{\lambda} + \hat{\alpha}\hat{P} \quad . \tag{8}$$

Since the network is open it follows that each entry of the matrix $\hat{P}^n$ converges to 0 as $n \to \infty$. Thus, the matrix $I - \hat{P}$ is invertible and the equation (8) has a unique solution for a given $\hat{\lambda}$. Intuitively, the equation (8) is a balance equation in which $\hat{\alpha}_i$ is interpreted as the total long-run average arrival rate to node i when the system is stable. And, it is plausible

that the system will be stable if the traffic intensity is less than one at each node, i.e., if

$$\rho_i \equiv \hat{\alpha}_i/\mu_i < 1$$

for $i = 1, 2, \ldots, N$.

A major result to be reported in Reiman (1977) is a heavy traffic limit theorem for the time-dependent behavior of the queue-lengths vector process $H(\cdot)$ as $\rho_i \uparrow 1$ for each $i = 1, \ldots, N$. This result is for a sequence of networks of queues as described above, and generalizes the result (5) and other results from Iglehart and Whitt (1970b) for acyclic networks in heavy traffic. The limiting process obtained is a N-dimensional diffusion $D(\cdot)$ whose state space is the non-negative orthant. On the interior of the state space $D(\cdot)$ behaves as a Brownian motion in N-dimensions with specified drift coefficients and covariance matrix. The covariance matrix depends on the switching matrix and the underlying means and variances of the interarrival times in the external input streams and the service times at the various nodes.

For the boundary behavior of the diffusion $D(\cdot)$, there is instantaneous reflection on each boundary surface, the angle of reflection being constant for each boundary surface but different for each surface. The angles of reflection depend only upon the switching matrix. Other properties of the diffusion $D(\cdot)$, along with other results as well, are to be reported in Reiman (1977).

3.0 CLOSED SYSTEMS

In this section we discuss some approximation results for closed Markovian networks of queues. Our discussion is based on work of Iglehart and Lemoine (1973), (1974), Lureau (1974), and Iglehart and Lalchandani (1973). For closed Markovian systems some approximation results are available for the equilibrium distribution of the vector of queue-length processes at the various nodes, when the number of customers in the system and the number of servers at each node are large; and, in this same setting, some approximations for time-dependent behavior, via diffusion processes without boundaries, have been developed. The most general closed Markovian system to which available approximation results on equilibrium distributions can be applied is the model of Posner and Bernholtz (1968), while the most general closed Markovian system to which available results on diffusion approximations can be applied in the model of Gordon and Newell (1967). However, in order to illustrate clearly both the nature and the usefulness of the available results we will discuss them in the context of a "repairman model" which is a special case of the systems studied in [21] and [54].

3.1 A Repairman Model

The model which serves as the basis for our discussion throughout this section is a generalization of the classical repairman problem (cf. Barlow (1962)). The model consists of n operating units which are subject to failure according to an exponential distribution with parameter $\lambda > 0$. These operating units are backed up by m_n spare units which can be used to replace any of the operating units that fail. At most n units can be operating at a given time. Two types of failure are possible. With probability p a failure of

type one occurs and the failed unit requires service from repair facility 1 which operates as a standard s_n^1 - server queue with exponential service time distribution having parameter $\mu_1 > 0$. Similarly, with probability $q = 1 - p$ the failed unit goes to facility 2 which has s_n^2 servers and exponential service time distribution with parameter $\mu_2 > 0$. The flow of units in the system is shown in Figure 5. One interpretation of this model

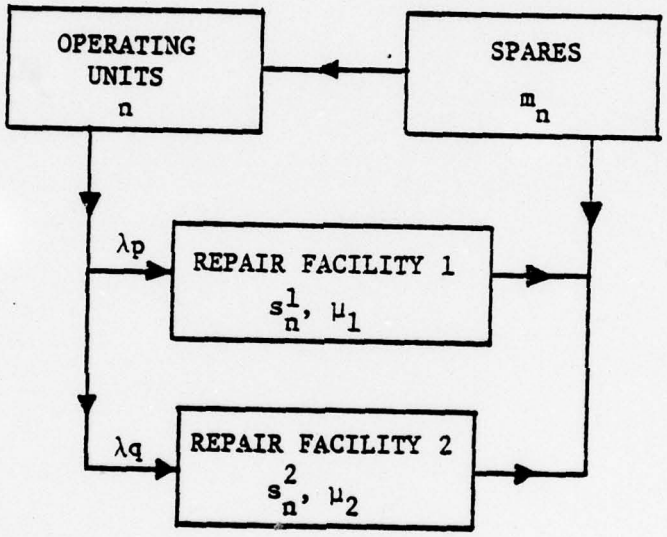

FIGURE 5 - REPAIRMAN MODEL

is that failures of type one are minor and can be repaired at a local repair facility, while failures of type two are major and require service at a central repair facility.

Let $X_n^k(t)$ denote the number of units waiting and undergoing repair at facility $k = 1, 2$ at time $t \geq 0$, and then let

$$X_n(t) = \begin{pmatrix} X_n^1(t) \\ X_n^2(t) \end{pmatrix} .$$

If $Y_n(t)$ denotes the number of operating units at time t, then $Y_n(t) = n - \left[X_n^1(t) + X_n^2(t) - m_n\right]^+$, where $x^+ = \max(x,0)$. Thus, the two-dimensional vector $X_n(t)$ is a meaningful description of the state of the system at time t. Since the repair and failure distributions are exponential the process $\{X_n(t), t \geq 0\}$ is an irreducible positive recurrent Markov chain with finite space $\Delta_n = \left\{\begin{pmatrix} i \\ j \end{pmatrix} : i \geq 0 \ , \ j \geq 0 \ , \ i + j \leq n + m_n\right\}$, depicted in Figure 6. We distinguish four regions of the state space and these are labeled A_n, B_n, C_n and D_n in Figure 6.

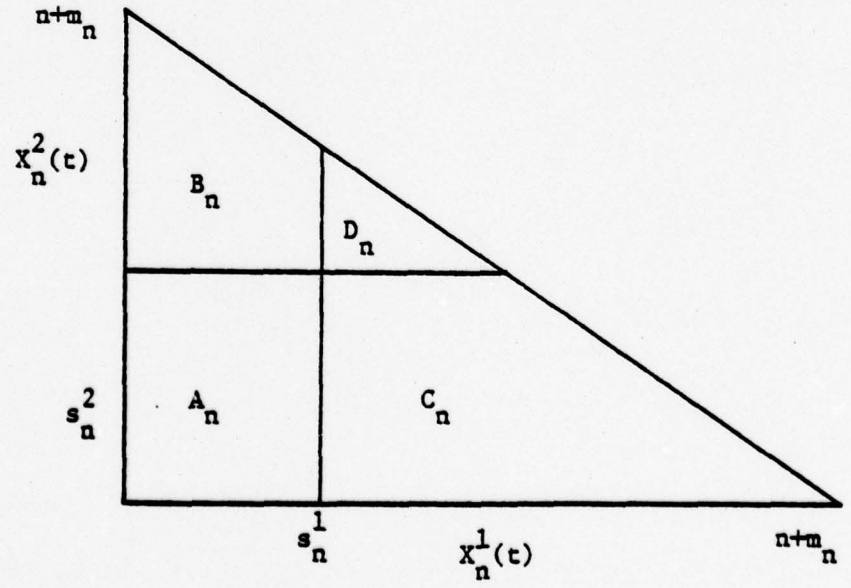

FIGURE 6 - STATE SPACE FOR $X_n(\cdot)$

The infinitesimal parameters of the process $X_n(\cdot)$ are developed as follows. For δ in Δ_n let $q(\delta)$ denote the total transition rate out of state δ and $q(\delta, \delta^*)$ the transition rate from state δ to state $\delta^* \neq \delta$. Let $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. If $X_n(t) = \delta$ then the next transition will be either to state $\delta + e_k$ (failed unit arrives at facility k) or to state $\delta - e_k$ (repaired unit leaves facility k) for $k = 1, 2$. Thus, for state δ we have

$$q(\delta) = q(\delta, \delta + e_1) + q(\delta, \delta + e_2) + q(\delta, \delta - e_1) + q(\delta, \delta - e_2)$$

where $q(\delta, \delta^*) = 0$ if δ^* is not in Δ_n. Table 1 gives the infinitesimal parameters for the four regions of the state space Δ_n depicted in Figure 6, where $x \wedge y = \min(x, y)$.

We want to approximate the behavior of the process $X_n(\cdot)$ when n is large. These approximations take two forms. In the first case we approximate the equilibrium or limiting behavior of the process; as we will see, this equilibrium distribution is difficult to work with, but simple approximations can be derived which yield much useful information. In the second case we approximate the process $\{X_n(t), t \geq 0\}$ itself by an appropriate diffusion process; these diffusion approximations are perhaps the only way of obtaining useful results on the time-dependent behavior of the system. The results we describe are derived under the assumption that the parameters s_n^1, s_n^2 and m_n grow linearly with n. Specifically, as $n \to \infty$ it is assumed that

$$s_n^k \sim n s_k \quad , \quad 0 < s_k < 1 \quad , \quad k = 1, 2 \qquad (9)$$

Parameters $\delta = \binom{i}{j}$ \ Region	A_n	B_n	C_n	D_n
$q(\delta, \delta+e_1)$	$\lambda p((n+m_n-i-j) \wedge n)$	$\lambda p((n+m_n-i-j) \wedge n)$	$\lambda p((n+m_n-i-j) \wedge n)$	$\lambda p((n+m_n-i-j) \wedge n)$
$q(\delta, \delta+e_2)$	$\lambda q((n+m_n-i-j) \wedge n)$	$\lambda q((n+m_n-i-j) \wedge n)$	$\lambda q((n+m_n-i-j) \wedge n)$	$\lambda q((n+m_n-i-j) \wedge n)$
$q(\delta, \delta-e_1)$	$i\mu_1$	$i\mu_1$	$s_n^1\mu_1$	$s_n^1\mu_1$
$q(\delta, \delta-e_2)$	$j\mu_2$	$s_n^2\mu_2$	$j\mu_2$	$s_n^2\mu_2$

TABLE 1: Infinitesimal Parameters for $\{X_n(t), t \geq 0\}$.

and

$$m_n \sim n m \ , \quad m > 0 \ . \tag{10}$$

The results obtained are in terms of the seven independent parameters in the model besides n : λ , ρ, s_1, μ_1, s_2, μ_2, m .

3.2 Approximating the Equilibrium Distribution

Let us denote by $\left\{ \pi_{ij}^{(n)} \ , \ \begin{pmatrix} i \\ j \end{pmatrix} \in \Delta_n \right\}$ the equilibrium on limiting distribution of the process $X_n(\cdot)$. This equilibrium distribution can be obtained from the results of [21], but since $X_n(\cdot)$ is also a reversible competition process, see Iglehart (1964), a more straightforward description of this distribution is possible. Accordingly, let $\gamma_{00}^{(n)} = 1$, and for $i + j > 0$ let $\gamma_{ij}^{(n)}$ be a ratio of products of infinitesimal parameters defined as follows. The numerator of $\gamma_{ij}^{(n)}$ is the product of parameters $q(\delta, \delta^*)$ for states δ and δ^* along the path (in the plane) from state $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ to state $\begin{pmatrix} i \\ j \end{pmatrix}$ which proceeds horizontally from state $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ to state $\begin{pmatrix} i \\ 0 \end{pmatrix}$ and then vertically from state $\begin{pmatrix} i \\ 0 \end{pmatrix}$ to state $\begin{pmatrix} i \\ j \end{pmatrix}$. Similarly, the denominator of $\gamma_{ij}^{(n)}$ is a product of parameters $q(\delta, \delta^*)$ for states δ and δ^* along the same path from $\begin{pmatrix} i \\ j \end{pmatrix}$ to $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ but in the opposite direction. Letting

$$G_n = \sum_{\Delta_n} \gamma_{ij}^{(n)} \ ,$$

we then have

$$\pi_{ij}^{(n)} = \lim_{t \to \infty} P\left\{X_n(t) = \begin{pmatrix} i \\ j \end{pmatrix}\right\} = \gamma_{ij}^{(n)}/G_n \qquad (11)$$

for each $\begin{pmatrix} i \\ j \end{pmatrix}$ in Δ_n .

It is rather obvious that for large n the probability distribution defined by (11) is tedious to calculate and, more importantly, difficult to use in drawing qualitative conclusions about overall system behavior. It is helpful, therefore, to have simple and useful approximations for this distribution. Such approximations were obtained in Iglehart and Lemoine (1973), (1974). And, some of these results were refined and extended in Lureau (1974), the extensions including approximations for the more general closed network model of Posner and Bernholtz (1968).

We now give a sample of the results from [30], [31] and [45] for the model of Figure 5

In what follows, let $X_n = \begin{pmatrix} X_n^1 \\ X_n^2 \end{pmatrix}$ be a random vector having the distribution defined by (11). We can thus regard X_n^k as the number of units at repair facility k under equilibrium conditions. Then, let $Y_n = n - [X_n^1 + X_n^2 - m_n]^+$ denote the number of operating units under equilibrium conditions. Let $a_1 = \lambda p/\mu_1$, $a_2 = \lambda q/\mu_2$ and $\alpha = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$. Finally, let $\mathcal{P}(\kappa)$ denote a Poisson random variable with parameter κ and $N(M, \Sigma)$ a random vector having a bivariate normal distribution with mean vector M and covariance matrix Σ .

The first result we give is perhaps the most favorable one from the standpoint of an operational system. Both repair facilities are operating in light traffic and an ample number of spares are available.

If $a_1 < s_1$, $a_2 < s_2$, and $s_1 + s_2 \leq m$, then

$$\pi_{ij}^{(n)} = \left[1 - o(n^{-1/2})\right] P\left\{\mathcal{P}(n a_1) = i\right\} P\left\{\mathcal{P}(n a_2) = j\right\}$$

for $\binom{i}{j}$ in $\Delta_n' = \left\{\binom{u}{v} : 0 \leq u \leq s_n^1 , 0 \leq v \leq s_n^2 , u + v \leq m_n\right\}$,

and the terms $o(n^{-1/2})$ are uniform over Δ_n' ;

$$\lim_{n\to\infty} P\left\{X_n \in \Delta_n'\right\} = 1 ;$$

$$(X_n - n\alpha)/n^{1/2} \Longrightarrow N(0, \Gamma) ,$$

where

$$\Gamma = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix} ;$$

and

$$\lim_{n\to\infty} P\left\{Y_n = n\right\} = 1 .$$

Thus, if we have a light traffic condition at each facility ($a_1 < s_1$ and $a_2 < s_2$) and an ample supply of spares ($a_1 + a_2 < m$), then with high probability no queues form at either facility and n units are in operation. Note that X_n^1 and X_n^2 are asymptotically independent, so that the two facility model in this case behaves exactly like two independent one facility models. Moreover, X_n^k has a distribution which is approximately normal with mean $n a_k$ and variance $n a_k$ for $k = 1, 2$.

The next result we present describes a situation which is far less satisfactory from an operational standpoint. In the case we consider, facility 1(2) is in heavy (light) traffic. Let $\nu = (1 + a_2)/a_1$, $\gamma = s_1/a_1$, and $\alpha^* = \begin{pmatrix} 1 + m - s_1\nu \\ \gamma \end{pmatrix}$.

<u>If</u> $a_1 > s_1$, $a_2 < s_2$, $s_1\nu < 1$, <u>and</u> $s_1 + s_2 \leq m$, <u>then</u>

$$\pi_{ij}^{(n)} = \left[1 - o(n^{-1/2})\right] P\left\{\mathcal{P}(n\gamma) = n + m_n - i - j\right\} \times P\left\{\mathcal{P}(n\gamma a_2) = j\right\}$$

<u>for</u> $\begin{pmatrix} i \\ j \end{pmatrix}$ <u>in</u> $\Delta_n'' = \left\{\begin{pmatrix} u \\ v \end{pmatrix} : u \geq m_n,\ 0 \leq v \leq s_n^2,\ u + v \leq n + m_n\right\}$,

<u>and the terms</u> $o(n^{-1/2})$ <u>are uniform over</u> Δ_n'' ; <u>and</u>

$$\lim_{n\to\infty} P\left\{X_n \in \Delta_n''\right\} = 1 .$$

In addition, if $|n^{-1}s_n^1 - s_1| = o(n^{-1/2})$ and $|n^{-1}m_n - m| = o(n^{-1/2})$, **then**

$$(X_n - n\alpha^*)/n^{-1/2} \Longrightarrow N(0, \Lambda) \quad ,$$

where

$$\Lambda = \begin{pmatrix} s_1\nu & -\gamma a_2 \\ -\gamma a_2 & \gamma a_2 \end{pmatrix} \quad ;$$

and

$$(Y_n - n\gamma)/(n\gamma)^{1/2} \Longrightarrow N(0, 1) \quad .$$

The above result describes a situation in which facility 1 is saturated but no queue forms at facility 2. The number of operating units fluctuates about $n s_1/a_1$, and with high probability fewer than n units are operating regardless of how many spares are provided, as long as $s_1 + s_2 \leq m$. Thus, we see that spares are of no help in alleviating the heavy traffic condition at facility 1; adding more spares only increases the congestion at facility 1 without producing more operating units. We remark that the requirement on the rate of convergence of $n^{-1}s_n^1$ to s , and of $n^{-1}m_n$ to m is not severe; it holds, for instance, in the reasonable case where $s_n^1 = [n s_1]$ and $m_n = [n m]$, the symbol $[\cdot]$ denoting integer part.

In addition to the two results presented above, a comprehensive set of cases for the model of Figure 5 is considered in [30], [31] and [45]; this includes the interesting case where both facilities are in light traffic but fewer than n units are operating, a circumstance in which providing more spares does result in having more units operational.

Thus, we see that while the model of Figure 5 does have an unwieldy equilibrium distribution, this distribution does admit simple approximations which provide useful information about system behavior.

3.3 Diffusion Approximations

The time-dependent behavior of the Markov chain $\{X_n(t), t \geq 0\}$ is virtually impossible to analyze in a direct manner. It does turn out, however, that as $n \to \infty$ the process $X_n(\cdot)$ with appropriate time and state scales does converge in distribution (or weakly) to a limiting diffusion process. The particular diffusion process of interest here is the bivariate Ornstein-Uhlenbeck process (b.O.U). The diffusion approximations we will describe follow from more general results in Iglehart and Lalchandani (1973) on the convergence of Markov chains in discrete-time and in continuous-time to the multi-variate Ornstein-Uhlenbeck process. These diffusion results in Iglehart and Lalchandani (1973) were derived by heuristic arguments, but using methods to be reported in Varadhan (1977), the results can now be rigorously justified. Moreover, these results can be applied to obtain diffusion approximations for the network model of Gordon and Newell (1967).

Before giving a sample of the approximation results for the model of Figure 5, we provide a brief description of the b.O.U. process; for further details see Schach(1971), Iglehart and Lalchandani (1973), and Arnold (1974). If we

say that $\{Y(t),\ t \geq 0\}$ is a b.O.U. process we mean that $Y(\cdot)$ is a two-dimensional Markov process having continuous sample paths and stationary transition probabilities, and for which there are real 2×2 matrices A and B, with A symmetric and positive definite, such that

$$\mu(t) \equiv E\{Y(t)\} = e^{-Bt} E\{Y(0)\} \quad ,$$

$$\Sigma(t) \equiv E\left\{[Y(t) - \mu(t)][Y(t) - \mu(t)]'\right\} = \int_0^t e^{-Bz} A e^{-B'z} dz \quad ,$$

and the conditional distribution of $Y(t + s)$ given $Y(s)$ is the bivariate normal

$$N\left(e^{-Bt} Y(s) \ , \ \Sigma(t)\right)$$

for all $t, s \geq 0$; the symbol $'$ denotes transpose and all vectors are taken to be column vectors. If the eigenvalues of the matrix B have strictly positive real parts, then $e^{-Bt} \to 0$ as $t \to \infty$, so that $\mu(t) \to 0$ as $t \to \infty$ and

$$C \equiv \lim_{t \to \infty} \Sigma(t) = \int_0^\infty e^{-Bz} A e^{-B'z} dz \quad ;$$

the matrix C is symmetric and positive definite and is the unique solution of the matrix equation

$$BD + DB' = A \quad .$$

Moreover, in this case

$$\Sigma(t) = C - e^{-Bt} C e^{-Bt}$$

and $Y(t)$ converges in distribution (as $t \to \infty$) to the bivariate normal $N(0, C)$. In the examples we consider, the matrix B will indeed have eigenvalues with strictly positive real parts.

The matrices B and A are related to the infinitesimal mean and the infinitesimal covariance of the process $Y(\cdot)$ in the sense that for any vector y and for any $t \geq 0$

$$\lim_{h \to 0} h^{-1} E\{Y(t + h) - Y(t) \mid Y(t) = y\} = -By$$

and

$$\lim_{h \to 0} h^{-1} E\left\{[Y(t + h) - Y(t)][Y(t + h) - Y(t)]' \mid Y(t) = y\right\} = A \quad .$$

In order to obtain a diffusion limit for the model of Figure 5, under the assumptions (9) and (10), we form the sequence of processes $\{Y_n(\cdot)$, $n = 1, 2, \ldots\}$ where

$$Y_n(t) = \frac{X_n(t) - ng}{n^{1/2}}$$

for $t \geq 0$ and g is a fixed two-dimensional vector selected so that the infinitesimal mean and covariance of the process $\{Y_n(t),\ t \geq 0\}$ converge as $n \to \infty$ to those of appropriate b.O.U. process $\{Y(t),\ t \geq 0\}$. We then approximate the behavior of $X_n(t)$ by that of $n^{1/2}Y(t) + n g$ when n is large. The vector $n g$ should be an "equilibrium point" of the Markov chain $X_n(\cdot)$ in that the chain should drift toward $n g$. In the notation of Section 3.1, the vector $n g$ lies within the triangle defined by Δ_n when n is large, and satisfies asymptotically the balance equations

$$q\left([ng], [ng] + e_1\right) = q\left([ng], [ng] - e_1\right)$$

and

$$q\left([ng], [ng] + e_2\right) = q\left([ng], [ng] - e_2\right)$$

as $n \to \infty$, where $[ng] = \begin{pmatrix} [ng_1] \\ [ng_2] \end{pmatrix}$ for $g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$.

We now give the diffusion analog for the first result of Section 3.2; we refer back to that result for the definition of various quantities which appear below.

<u>Let</u> $g = \alpha$ <u>and suppose that</u> $X_n(0) = [ng]$. If $a_1 < s_1$, $a_2 < s_2$, and $s_1 + s_2 \leq m$, <u>then for every</u> $t \geq 0$

$$Y_n(t) \Longrightarrow Y(t) \qquad (12)$$

<u>as $n \to \infty$, where $\{Y(t),\ t \geq 0\}$ is a b.O.U. process with</u>

$$A = \begin{bmatrix} 2\lambda p & 0 \\ 0 & 2\lambda q \end{bmatrix}$$

<u>and</u>

$$B = \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix}$$

<u>and</u> $C = \Gamma$.

Given the relationship among the basic system parameters in the above, the vector ng lies in the region A_n of Figure 6 for n large, and thus so does $X_n(0)$. From (12) we see that $X_n(t)$ (or $n^{1/2} Y(t) + ng$) has fluctuations about ng of the order $n^{1/2}$; but, the distance from the boundaries of region A_n to the point ng is of the order n . Thus, the process $\{X_n(t),\ t \geq 0\}$ never leaves the region A_n with any appreciable probability. Since $na_1 + na_2 < m_n$, we therefore see that for large n there are n units in operation at time t with high probability. Since the matrix C has zero off-diagonal elements, the component processes $\{X_n^1(t),\ t \geq 0\}$

and $\{X_n^2(t),\ t \geq 0\}$ are asymptotically independent. Each component process, appropriately normalized, converges to a univariate Ornstein-Uhlenbeck process. Hence, for large n, $X_n^1(t)$ has a distribution which is approximately $N\left(n a_1 e^{-\mu_1 t},\ n a_1(1 - e^{-2\mu_1 t})\right)$ and $X_n^2(t)$ has a distribution which is approximately normal $N\left(n a_2 e^{-\mu_2 t},\ n a_2(1 - e^{-2\mu_2 t})\right)$. Letting $t \to \infty$ in (12), we find that $(X_n - ng)/n^{1/2} \Longrightarrow N(0, C)$, in agreement with the first result of Section 3.2.

The next result we present is the diffusion analog of the second result from Section 3.2. We refer to that result for the definition of various quantities which appear below.

<u>Let</u> $g = \alpha^*$ <u>and suppose that</u> $X_n(0) = [ng]$. If $a_1 > s_1$, $a_2 < s_2$, $s_1\nu < 1$, <u>and</u> $s_1 + s_2 \leq m$, <u>then for every</u> $t \geq 0$

$$Y_n(t) \Longrightarrow Y^*(t) \tag{13}$$

<u>as</u> $n \to \infty$, <u>where</u> $\{Y^*(t),\ t \geq 0\}$ <u>is a b.O.U. process with</u>

$$A = \begin{bmatrix} 2\mu_1 s_1 & 0 \\ 0 & 2\mu_1 s_1 q/p \end{bmatrix}$$

<u>and</u>

$$B = \begin{bmatrix} \lambda p & \lambda p \\ \lambda p & \mu_2 + \lambda q \end{bmatrix}$$

and $C = \Lambda$.

Given the relationship among the basic system parameters in the above, the vector $n\,g$ lies in the region C_n of Figure 6 when n is large. Thus, the process $\{X_n(t),\ t \geq 0\}$ never leaves C_n with any appreciable probability. Facility 1 is saturated while facility 2 is stable, and far fewer than n units are operating. Additional spares are of no help in alleviating congestion; in fact, they only exacerbate an already undesirable situation. Letting $t \to \infty$ in (13) we see that $(X_n - n\,g)/n^{1/2} \Longrightarrow N(0, C)$, in agreement with the second result of Section 3.2. Thus, for n large and t large, the number of operating units has a distribution which is approximately $N(n\,s_1/a_1,\ n\,s_1/a_1)$.

Finally, we remark that the diffusion approximations given here form a small sample of the results from the paper of Iglehart and Lalchandani (1973). In addition to the system of Figure 5, the paper also provides diffusion limits for other repair systems, as well as the general results on weak convergence of Markov chains mentioned above.

4.0 SOME OPEN PROBLEMS

In this section we cite some open problems in the area of approximation results for networks of queues. Given the complex nature of queueing networks and the available approximation results which have been discussed in Sections 2 and 3, one can compose a rather long list of open problems in the area. There are, however, a certain few problems which seem to merit careful study, and we will now discuss these in separate numbered paragraphs.

1. Consider a closed network model of the sort introduced by Gordon and Newell (1967) but with arbitrarily distributed service times. There are no general results available for such systems, save some equilibrium results and some occupation-time results for a few special cases. It would therefore be helpful to develop diffusion approximations for closed systems with arbitrarily distributed service times in the spirit of those obtained by Iglehart and Lalchandani (1973) for closed Markovian systems.

2. The limiting diffusion approximations which have been developed for open networks involve complicated functionals of multi-dimensional Brownian motion. Except in the case of single-node systems, however, there are as yet no computationally tractable approximations of general applicability which have been extracted from these diffusion limits for multi-node systems. In particular, first-passage or occupation-time distributions are as yet unavailable for these limiting diffusion processes. Thus, further work is necessary in order to make existing results more useful for answering a variety of questions which arise in the contexts of system design and system control.

3. With few exceptions, the queueing theory literature is concerned with systems whose arrival rates (distributions) and service rates (distributions)

do not vary with time. It would be very interesting, for example, to develop a diffusion limit for an open network model of the sort introduced by Jackson (1957) but for which the external input streams are non-stationary Poisson processes having rate functions which are periodic (as functions of time).

ACKNOWLEDGEMENT. We are grateful to J. M. Harrison and M. Reiman for extensive discussions on their work prior to its publication.

REFERENCES

[1] ARNOLD, L. (1974). Stochastic Differential Equations. Wiley, New York.

[2] BARLOW, R. E. (1962). Repairman Problems. Studies in Applied Probability and Management Science. Eds. K. Arrow, S. Karlin, and H. Scarf. Stanford University Press, Stanford, California. 18-33.

[3] BILLINGSLEY, P. (1968). Convergence of Probability Measures. Wiley, New York.

[4] BREIMAN, L. (1968). Probability. Addison-Wesley, Reading, Mass.

[5] COX, D. R. and MILLER, H. D. (1965). The Theory of Stochastic Processes. Metheun, London.

[6] CRANE, M. (1971). Limit Theorems for Queues in Transportation Systems. Technical Report No. 16, Department of Operations Research, Stanford University.

[7] CRANE M. (1973). Queues in Transportation Systems, I: A Markovian System. J. Appl Prob. 10, 630-643.

[8] CRANE, M. (1974a). Queues in Transportation Systems, II: An Independently Dispatched System. J. Appl. Prob. 11, 145-158.

[9] CRANE, M. (1974b). Multi-Server Assembly Queues. J. Appl. Prob. 11, 629-632.

[10] DYNKIN, E. B. (1965). Markov Processes, Vols. I and II. Springer-Verlag, Berlin.

[11] FELLER, W. (1971). An Introduction to Probability Theory and Its Applications, Vol II. 2nd Ed. Wiley, New York.

[12] FREEDMAN, D. (1971). Brownian Motion and Diffusion. Holden-Day, San Francisco.

[13] FRIEDMAN, A. (1975). Stochastic Differential Equations and Applications, Vol 1. Academic Press, New York.

[14] FRIEDMAN, A. (1976). Stochastic Differential Equations and Applications, Vol. 2. Academic Press, New York.

[15] GAVER, D. P. (1968). Diffusion Approximations and Models for Certain Congestion Problems. J. Appl. Prob. 5, 607-623.

[16] GAVER, D. P., LEHOCZKY, J. P., and PERLAS, M. (1975). Service Systems with Transitory Demand. Studies in the Management Sciences, Vol. I, Logistics. North-Holland, TIMS. 21-34.

[17] GAVER, D. P. and LEHOCZKY, J. P. (1976a). Gaussian Approximations to Service Problems: A Communication System Example. J. Appl. Prob. 13, 768-780.

[18] GAVER, D. P. and LEHOCZKY, J. P. (1976b). A Diffusion Approximation Model for a Communication System Allowing Message Interference. Preprint.

[19] GIKHMAN, I. I. and SKOROKHOD, A. V. (1969). An Introduction to the Theory of Random Processes. W. B. Saunders, Philadelphia.

[20] GIKHMAN, I. I. and SKOROKHOD, A. V. (1972). Stochastic Differential Equations. Springer-Verlag, Berlin.

[21] GORDON, W. J. and NEWELL, G. F. (1967). Closed Queueing Systems with Exponential Servers. Operations Res. 15, 254-265.

[22] HARRISON, J. M. (1973a). Assembly-Like Queues. J. Appl. Prob. 10, 354-367.

[23] HARRISON, J. M. (1973b). The Heavy Traffic Approximation for Single Server Queues in Series. J. Appl. Prob. 10, 613-629.

[24] HARRISON, J. M. (1977): Tandem Queues in Heavy Traffic. Forthcoming.

[25] IGLEHART, D. L. (1964). Reversible Competition Processes. Z. Wahrscheinlichkeithstheorie verw. Gebiete. 2, 314-331.

[26] IGLEHART, D. L. and WHITT, W. (1970a). Multiple Channel Queues in Heavy Traffic. I. Adv. Appl. Prob. 2, 150-177.

[27] IGLEHART, D. L. and WHITT, W. (1970b). Multiple Channel Queues in Heavy Traffic. II: Sequences, Networks and Batches. Adv. Appl. Prob. 2, 355-369.

[28] IGLEHART, D. L. (1973). Weak Convergence in Queueing Theory. Adv. Appl. Prob. 5, 570-594.

[29] IGLEHART, D. L. and LALCHANDANI, A. P. (1973). Diffusion Approximations for Complex Repair Systems. Technical Report No. 266-12, Control Analysis Corporation, Palo Alto, California.

[30] IGLEHART, D. L. and LEMOINE, A. J. (1973). Approximations for the Repairman Problem with Two Repair Facilities, I: No Spares, Adv. Appl. Prob. 5, 595-613.

[31] IGLEHART, D. L. and LEMOINE, A. J. (1974). Approximations for the Repairman Problem with Two Repair Facilities, II: Spares. Adv. Appl. Prob. 6, 147-158.

[32] ITÔ, K. (1961). Lectures on Stochastic Processes. Tata Institute of Fundamental Research, Bombay.

[33] ITÔ, K. and McKEAN, H. P. (1965). Diffusion Processes and Their Sample Paths. Springer-Verlag, Berlin.

[34] JACKSON, J. R. (1957). Networks of Waiting Lines. Operations Res. 5, 518-521.

[35] KARLIN, S. and TAYLOR, H. (1975). A First Course in Stochastic Processes. 2nd Ed. Academic Press, New York.

[36] KENNEDY, D. P. (1972a). Rates of Convergence for Queues in Heavy Traffic. I. Adv. Appl. Prob. 4, 357-381.

[37] KENNEDY, D. P. (1972b). Rates of Convergence for Queues in Heavy Traffic. II. Sequences of Queueing Systems. Adv. Appl. Prob. 4, 382-391.

[38] KINGMAN, J. F. C. (1961). The Single Server Queue in Heavy Traffic. Proc. Camb. Phil. Soc. 57, 902-904.

[39] KINGMAN, J. F. C. (1962). On Queues in Heavy Traffic. J. R. Statist. Soc. B 24, 383-392.

[40] KINGMAN, J. F. C. (1965). The Heavy Traffic Approximation in the Theory of Queues. Proc. Symposium on Congestion Theory. Eds. W. Smith and W. Wilkinson, Univ. of North Carolina Press, Chapel Hill. 137-159.

[41] KOBAYASHI, H. (1974a). Application of the Diffusion Approximation to Queueing Networks, I. Equilibrium Queue Distributions. J. Assoc. Comput. Mach. 21, 316-328.

[42] KOBAYASHI, H. (1974b). Application of the Diffusion Approximation to Queueing Networks, II. Nonequilibrium Distributions and Computer Modeling. J. Assoc. Comput. Mach. 21, 459-469.

[43] KÖLLERSTRÖM, J. (1974). Heavy Traffic Theory for Queues with Several Servers. I. J. Appl. Prob. 11, 544-552.

[44] LEMOINE, A. J. (1977). Networks of Queues - Equilibrium Analysis. Technical Report No. 19-1, Control Analysis Corporation, Palo Alto, California.

[45] LUREAU, F. (1974). A Queueing Theoretic Analysis of Logistics Repair Models with Spare Parts. Technical Report No. 55, Department of Operations Research, Stanford University.

[46] McNEIL, D. R. (1973). Diffusion Limits for Congestion Models. J. Appl. Prob. 10, 368-375.

[47] McNEIL, D. R. and SCHACH, S. (1973). Central Limit Analogues for Markov Population Processes. J. R. Statist. Soc. B. 35, 1-23.

[48] NEWELL, G. F. (1968). Queues with Time Dependent Arrival Rates. I, II, III. J. Appl. Prob. 5, 436-451, 579-590, 591-606.

[49] NEWELL, G. F. (1971). Applications of Queueing Theory. Chapman and Hall, London.

[50] NEWELL, G. F. (1973). Approximate Stochastic Behavior of n-Server Service Systems with Large n. Lecture Notes in Economics and Mathematical Systems No. 87. Springer-Verlag, Berlin.

[51] NEWELL, G. F. (1975). Approximate Behavior of Tandem Queues. Special Report, Institute of Transportation and Traffic Engineering, University of California, Berkeley.

[52] NEWELL, G. F. (1977). Approximate Behavior of Tandem Queues. II. Forthcoming.

[53] PARTHASARATHY, K. R. (1967). Probability Measures on Metric Spaces. Academic Press, New York.

[54] POSNER, M. and BERNHOLTZ, B. (1968). Closed Finite Queueing Networks with Time Lags. Operations Res. 5, 962-976.

[55] REIMAN, M. (1977). Queueing Networks in Heavy Traffic. Forthcoming.

[56] SCHACH, S. (1971). Weak Convergence Results for a Class of Multivariate Markov Processes. Ann. Math. Statist. 42, 451-565.

[57] VARADHAN, S. R. S. (1968). Stochastic Processes. Courant Institute of Mathematical Sciences, New York.

[58] VARADHAN, S. R. S. (1977). Diffusion Processes and Their Applications. Forthcoming.

[59] WHITT, W. (1968). Weak Convergence Theorems for Queues in Heavy Traffic. Technical Report No. 2, Department of Operations Research, Stanford University.

[60] WHITT, W. (1974). Heavy Traffic Limit Theorems for Queues: A Survey. Mathematical Methods in Queueing Theory. Lecture Notes in Economics and Mathematical Systems No. 98. Springer-Verlag, Berlin. 307-350.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (14) TR- Technical Report No. 19-2	2. GOVT ACCESSION NO. (9)
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) (6) NETWORKS OF QUEUES - APPROXIMATION RESULTS.	5. TYPE OF REPORT & PERIOD COVERED Final Technical Report July 1976 - June 1977
6. PERFORMING ORG. REPORT NUMBER N/A	
7. AUTHOR(s) (10) Austin J. Lemoine	8. CONTRACT OR GRANT NUMBER(s) (15) N00014-76-C-0919
9. PERFORMING ORGANIZATION NAME AND ADDRESS Systems Control, Inc. 1801 Page Mill Road Palo Alto, CA 94304	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-365
11. CONTROLLING OFFICE NAME AND ADDRESS (11) Office of Naval Research Code 436 Arlington, VA 22217	12. REPORT DATE May 16, 1977 13. NUMBER OF PAGES 50
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) (12) 55p. (same)	15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

(same)

18. SUPPLEMENTARY NOTES

ONR Project Monitor
Dr. Bruce J. McDonald

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Networks of Queues, approximation results, Open networks, Closed networks, Queue-lengths process, Waiting time process, Heavy traffic, Brownian motion, Diffusion processes

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is a sequel to our earlier review paper Lemoine (1977) on the equilibrium analysis of networks of queues. In this report we review some approximation results for networks of queues. Detailed discussion is limited to results which can be rigorously justified. In addition, we call attention to some important open problems. The bibliography includes references for results derived in a heuristic or informal fashion and includes background material on diffusion processes.

389 333